心理科学进展 2020, Vol. 28, No. 5, 800–809
Advances in Psychological Science DOI: 10.3724/SP.J.1042.2020.00800

信任的认知神经网络模型*

陈 瀛 徐敏霞 汪新建
(南开大学周恩来政府管理学院社会心理学系，天津 300350)

摘 要 信任是指一方在基于对另一方行为期望的基础上愿意冒一定的风险，以期在将来得到积极结果的心理过程。近年，认知神经取向的研究对信任行为引起的特定脑区激活进行了考察，却忽略了大规模脑网络在信任过程中的整体作用。在总结前人研究的基础上提出信任的认知神经网络模型，并从认知神经网络视角对信任行为进行解释和整合。在模型中，信任行为是动力系统、情感系统和认知系统相互作用的结果，并分别与奖励网络、显著网络、中央执行网络和默认网络等神经网络激活有关。此外，模型还强调信任行为的反馈机制，以此构成完整的建构模型。模型阐明了心理系统与中枢神经网络之间的对应关系，从认知神经角度解释了信任行为发生的心理机制和神经基础。

关键词 信任；认知神经网络；动力系统；情感系统；认知系统
分类号 B845; B849:C91

1 引言

信任是指一方在基于对另一方行为期望的基础上愿意冒一定的风险，以期在将来得到积极结果的心理过程(Krueger & Meyer-Lindenberg, 2018)。信任遍布于社会生活的方方面面，被称为“社会的润滑剂” (Arrow, 1972)，在人际和群际层面都发挥着重要作用。人际层面，个体间的交往与合作往往建立在信任的基础之上，如陌生人之间关系的形成与维系(Clark & Eisenstein, 2013)；群际层面，人类社会互动也必须建立在信任的基础之上，否则就可能出现认知偏差乃至误解，如种族偏见的形成(Stets & Fares, 2019)。因此，信任是维系社会稳定和谐的重要因素(Rothstein & Uslaner, 2005)。然而，信任关系又具有不稳定性，这就带来了社会两难问题：如果人们都轻信他人，就会导致社会中的骗子越来越多；如果人们都不信任对方，又无法进行正常的人际交往(Irwin, Edwards, & Tamburello, 2015)。

信任在社会生活中的重要作用引起了国内外学者们的重视，经济学、心理学和神经科学等不同领域的学者们对信任问题从理论和实践角度进行了大量研究，并取得了丰硕成果。基于博弈论提出的信任博弈游戏(trust game, TG)是研究信任的重要方法之一，在经典的信任博弈范式中，两名被试分别扮演商业投资中的委托人(A)和受托人(B)，在实验的开始阶段双方各得到 10 点出场费，A 需要决定拿出多少钱(X)给 B，作为投资收益它会变成原来的 3 倍(3X)；当 B 得到这些钱后需要决定返还给 A 多少钱，记做 Y。因此，A 在投资中的总收益为 10−X+Y，而 B 在投资中的总收益为 10+3X−Y。根据经济人假设(Smith & Skinner, 2000)，为了追求利益最大化，B 的最佳策略是保留所有投资的钱(不返还)，从而保证得到最大收益 10+3X；如果 A 知道了 B 所采取的策略，那么他就不会进行投资，即 X=0。这就是博弈论中对信任博弈游戏进行预测所达到的纳什平衡(Nash equilibrium) (Fang & Wu, 2019)。而实际研究中得到的结果却与预测结果大相径庭，大量重复研究发现委托人投资金额大约是总金额的一半，而受托人往往也会返还给委托人略小于总金额一半的

收稿日期：2019-09-20
* 教育部哲学社会科学研究重大课题攻关项目“医患信任关系建设的社会心理机制研究” (15JZD030)、天津市社会科学规划项目“天津市心理健康服务体系的实务模式探索” (TJJX18-001)和南开大学博士研究生科研创新基金资助。
通信作者：汪新建, E-mail: wangxj@nankai.edu.cn

钱作为投资回报(Berg, Dickhaut, & McCabe, 1995)。通过一系列信任博弈游戏研究发现，人类在风险投资中存在着信任和互惠行为。信任博弈范式不但能够发挥实验室研究可量化和可重复性特点，还可以对实际货币行为进行模拟并探讨互惠信任中的基本特征，因此得到进一步完善(图 1)，并迅速应用到心理学研究之中。除信任博弈范式外，在信任领域研究中常见的范式还有囚徒困境(prisoner's dilemma)、议价博弈(bargain game)、重复信任博弈(repeated trust game)、三者信任博弈(three-player trust game)和蜈蚣博弈(centipede game)等，在各种实验研究及其变式中都给我们展示了这样一种社会困境：即一方是否会基于对未来奖励的预期而愿意冒可能会被背叛的风险，并选择相信另一方(Anderhub, Engelmann, & Güth, 2002; Engle-Warnick & Slonim, 2004)。

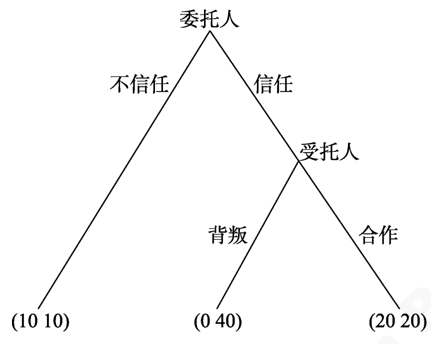


图 1　简化信任博弈游戏示意图
注：括号中数字分别为委托人和受托人的收益

在信任领域的诸多研究中，认知神经取向占有重要地位，该取向以神经科学为基础对信任的发生机制进行探讨，以期从神经层面对信任机制给出更客观的指标，并做出更科学的解释。如Tzieropoulos (2013)从认知神经角度对信任博弈游戏中的信任问题进行剖析并详细探讨了包括激素、基因、个体差异以及实验设置等因素对信任博弈结果的影响；Bellucci, Chernyak, Goodyear, Eickhoff 和 Krueger (2017)使用元分析方法对信任博弈游戏中的互惠行为进行深入研究，结果发现前脑岛、顶内沟、腹侧纹状体和背侧纹状体在信任和互惠的不同阶段各有激活；张宁、张雨青和吴坎坎(2011)总结了信任的不同模型，并指出信任行为与催产素水平有关，与背侧纹状体、前脑岛、伏隔核、尾状核等处理奖赏信息的脑区也有关；张蔚、张震、高宇、段华平和吴兴南(2016)总结了建立在信任博弈范式基础上的背叛厌恶理论、社会规范理论、道德规范理论和默认行为模型，并发现信任行为涉及的脑区有内侧前额叶、尾状核、杏仁核和脑岛等区域。

对以上研究进行总结可以发现，与信任有关的脑区可分为四个部分，它们分别是：(1)与奖励有关的脑区，如纹状体、伏隔核等; (2)与情绪有关的脑区，如杏仁核、脑岛等; (3)与逻辑推理和决策有关的脑区，如前额皮层等; (4)与社会功能有关的脑区，如顶内沟等。这些研究虽然从认知神经角度指出了与信任行为有关的因素和脑区，但信任作为人类的高级心理过程，其发生和发展过程都极为复杂，受到主观因素和客观因素等各种变量影响，因此在信任过程中大脑特定脑区的激活并非彼此孤立，而是以神经网络的形式进行激活。因此，从认知神经网络角度对信任进行探讨，有助于我们对信任发生和发展的神经机制有更深入的了解。

2　信任过程涉及的大规模脑网络

基于神经科学的观点(Bressler & Menon, 2010)，信任源于心理系统(动力系统、情感系统和认知系统)的相互作用，其中动力系统是指心理活动的内在力量或驱动力，是心理系统中负责寻求奖励并避免惩罚的结构；情感系统是指心理活动中涉及感受、情绪和情感反应(包括生理反应)的个性单元，个体对重要的外界信息加工时往往会伴随情绪性并产生对应的情感体验；认知系统是指心理活动中与思维、计划和行动相关的心理结构，通过外部感官收集信息并直接支配个体思维与行动。在人际信任中存在的不确定性驱使个体采取可靠的策略来准确评估对方的可信度，从而保证人际交往的正常进行。这些成分与心理系统中的动机、情感和认知相联系：对奖励的预期(动机)和对背叛危险性的评估(情感)共同作用产生不确定性，使人们对他人的信任具有脆弱性。为了减少不确定性，人们通常采用两种不同类型的有限理性(认知)，即经济理性(economic rationality)和社会理性(social rationality)。经济理性是指信任由外在激励驱动，即追求自身利益：当自身利益与集体利益一致时，人们更容易信任他人；社会理性是指信任由内在激励驱动：社会理性有助于人际关系的建立并重视群体的归属(Declerck, Boone,

& Emonds, 2013)。心理系统间的相互作用涉及大规模脑网络(large-scale brain networks)中的关键区域，他们的工作方式是以网络形式进行，并把若干特定脑区有机地联系在一起。所谓大规模脑网络是指分布在整个大脑中广泛互连的大脑皮层的集合，这些区域相互作用以执行各项心理功能。与采用基于任务的神经影像学研究相比，个体大规模脑网络系统中的静息态功能连接更能预测个体在信任行为中的差异(Bellucci, Hahn, Deshpande, & Krueger, 2019)。这些关键区域包括四个网络系统，分别是：奖励网络(reward network, RWN)、显著网络(salience network, SAN)、中央执行网络(central-executive network, CEN)和默认网络(default-mode network, DMN) (图 2)。其中，奖励网络作为信任的动力系统负责确定信任他人可能得到的奖励；显著网络作为信任的情感系统负责评估由于他人背叛所带来的风险并产生对风险的厌恶感；中央执行网络作为信任的认知控制系统负责决定在信任过程中采取基于情境的何种策略；默认网络作为信任的社会认知系统通过评估关系的可信度决定是否信任合作伙伴。

2.1 奖励网络：预期奖励

经济人假设认为人们不会轻信他人，除非信任他人能够给自己带来好处。现实生活中虽然不尽如此，但从信任的定义可以看出，个体之所以选择相信他人，是因为期望“在将来得到积极的结果”，而这个“积极的结果”正是个体信任他人的动力。如果一个人不寻求在将来得到某些好处，那将是一种“布施行为”，而非信任行为。个体往往需要评估信任他人可能得到的奖励，来做出是否信任的决定。该过程涉及的脑区与奖励有关，被称为奖励网络(Davidenko et al., 2018)。

奖励网络是信任动力系统的神经基础，负责对奖励进行预期，其中包括中脑边缘通路、中脑皮质通路和黑质纹状体通路。中脑边缘通路将中脑腹侧被盖区(VTA)连接到腹侧纹状体(vSTR)的伏隔核和嗅结节；中脑皮质通路指从 VTA 到前额皮层(PFC)，包括腹内侧前额皮层(vmPFC)；黑质纹状体通路指从中脑黑质(SN)到尾状核和背侧纹状体(dSTR)壳核(Ikemoto, 2010)。中脑边缘通路中的 vSTR 调节与奖励有关的个人动机并易化信任的建立过程：面对值得信赖的合作伙伴比面对不

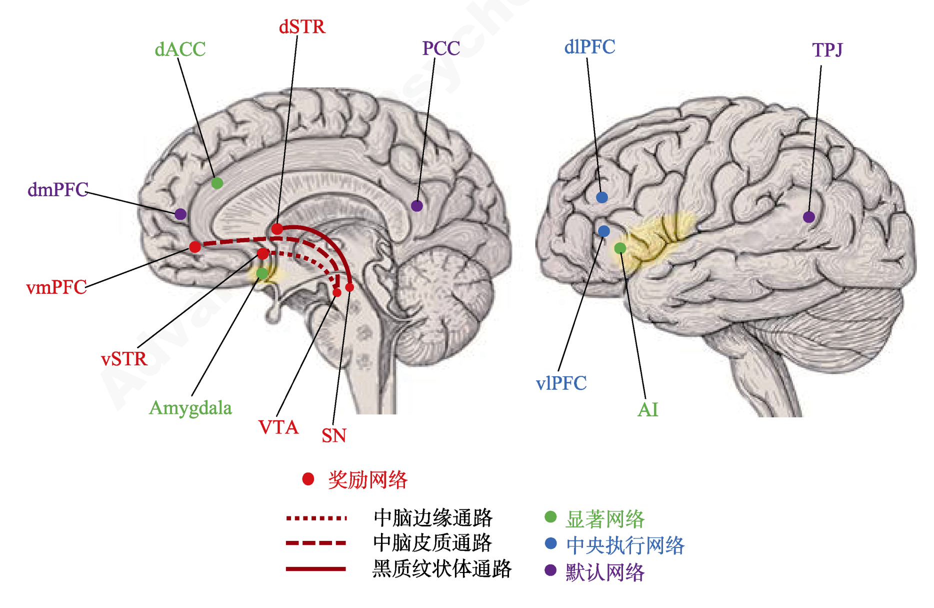


图 2　信任的认知神经网络

注：奖励网络包括中脑边缘通路、中脑皮质通路和黑质纹状体通路，涉及的脑区有中脑腹侧被盖区(VTA)、腹侧纹状体(vSTR)、腹内侧前额皮层(vmPFC)、黑质(SN)和背侧纹状体(dSTR)等；显著网络涉及的脑区有杏仁核(Amygdala)、前脑岛(AI)和背侧前扣带回(dACC)等；中央执行网络涉及的脑区有背外侧前额皮层(dlPFC)和腹外侧前额皮层(vlPFC)等；默认网络涉及的脑区有颞顶联合区(TPJ)、后扣带皮层(PCC)以及背内侧前额皮层(dmPFC)等。

值得信赖的合作伙伴时预期得到的奖励更高，此时负责调节奖励的 vSTR 激活程度更高，个体信任他人的动机增加，使个体更容易做出信任的决策(Delgado, Frank, & Phelps, 2005)。中脑皮质通路中的 vmPFC 与 vSTR 和杏仁核(amygdala)相连，被看作是理想的“神经整合器(neural integrator)”，它与编码刺激的预期奖励相关联，并负责将复杂且性质不同的奖励与主观价值相比较(Hare, Camerer, & Rangel, 2009)。社会情境中的异常决策往往与 vmPFC 损伤相关：当个体的 vmPFC 受损，往往不能正确编码刺激的性质，因此无法正确判断其能否带来奖励，也无法将其与个体已经形成的主观价值相比较，最终会导致盲目信任行为的增加(Moretto, Sellitto, & di Pellegrino, 2013)。黑质纹状体通路与运动产生有关，并影响计划(planning)和动作选择(action selection) (Joel & Weiner, 2000)。在重复信任博弈游戏中已证实 vSTR 激活对应游戏中决策阶段而 dSTR 激活对应游戏中的反馈阶段：在信任决策阶段 vSTR 激活，证明其在信任博弈过程中的决策阶段能够评估预期得到的奖励，当合作伙伴表现出值得信赖时会进行更多的投资；而在反馈阶段 dSTR 激活，即为了使投资人在接下来的重复博弈中有更多的投资，个体会选择给予相应的回报，表明其在信任博弈过程中影响计划的执行(Bellucci et al., 2017)。

2.2　显著网络：评价风险

信任行为从本质上是一种风险决策，因为当个体选择信任他人时可能得到积极的结果也可能得到消极的结果，所以做出选择时往往要冒一定的风险。当得到积极结果时会产生积极情绪，并促使个体在接下来的行为中更倾向信任对方；而得到消极结果时会产生消极情绪，使个体在接下来的行为中表现出不信任行为。由于信任行为具有风险性的特点，在不知道对方会做出何种行为的情况下，个体需要评价由他人背叛所带来的风险，而这种背叛的风险往往使人产生消极的情绪体验。该过程涉及的脑区与情绪(尤其是负性情绪)有关，被称为显著网络(Menon, 2015)。

显著网络是信任情感系统的神经基础，负责对风险进行评价，包括杏仁核、前脑岛(AI)和背侧前扣带回(dACC)等重要区域(Bressler & Menon, 2010)。杏仁核作为产生危险信号的脑区，能够编码情绪刺激并使个体更具社会警惕性。一项 fMRI 研究表明，当面对不值得信赖的面孔时杏仁核被激活，即信任度越低杏仁核激活程度越高(Engell, Haxby, & Todorov, 2007)。相比于正常被试，杏仁核受损的被试往往表现得更加“仁慈”：即使对方选择背叛，杏仁核受损的被试依然在信任博弈游戏中选择信任对方(Koscik & Tranel, 2011)。AI 负责对不可预测事件进行编码并产生主观厌恶感：当个体选择相信他人却因错信而导致损失时，AI 释放对背叛的厌恶信号(Namkung, Kim, & Sawa, 2017)。AI 损伤同样会引起非正常的信任行为：该类患者扮演委托人时，当对方表现出背叛行为，由于 AI 受损不能产生相应的厌恶信号，导致他们依然选择信任对方；而在扮演受托人时，他们往往不能对背叛产生正常的情绪体验，因此常常违反合作伙伴的信任，表现出更多的背叛行为。研究结果揭示出 AI 有助于在决策过程中识别风险和违反社会规范的行为(Belfi, Koscik, & Tranel, 2015)。dACC 作为边缘系统的重要组成部分与情绪和认知功能有关，负责对冲突进行监测以更好的适应社会(Bush, Luu, & Posner, 2000)，该区域不仅负责对恐惧和焦虑等负性情绪进行评估，还对情绪的表达具有调节作用(Etkin, Egner, & Kalisch, 2011)。当对方反复做出不值得信任的行为时，委托人的 dACC 激活状态更高(Fett, Gromann, Giampietro, Shergill, & Krabbendam, 2012)。

2.3　中央执行网络：选择策略

在做出信任决策前，人们往往需要对具体情境进行评估，最终做出决策。这一过程需要个体对所处情境进行综合判断，根据所有可得性线索，经过严谨的逻辑分析，最终确定使用何种策略进行信任决策，使得到的收益最大且损失最小。因此这一过程是基于计算、纯理性的思维过程。由于此过程需要大量精细的计算并在计算结果的基础上进行推理，最终做出决策，因此需要中央执行网络参与(Sherman et al., 2014)。

中央执行网络是信任认知控制系统的核心，负责对策略进行选择，包括背外侧前额皮层(dlPFC)和腹外侧前额皮层(vlPFC)等重要脑区(Miller & Cohen, 2001)。dlPFC 与眶额皮层、丘脑、基底神经节、海马等脑区相连，参与风险和道德决策，当个体必须做出分配有限资源的道德决定时 dlPFC 激活(Greene, Sommerville, Nystrom, Darley, & Cohen, 2001)。同样，当个体需要在两个不同的备

选方案中做出选择时，dlPFC 的激活会使个体做出更加公平的选择偏好，并抑制个人利益最大化的诱惑(Knoch & Fehr, 2007)。vlPFC 在运动抑制中发挥着重要作用：对于右利手个体而言，在更新动作计划时右后 vlPFC 激活；在不确定条件下进行决策时右中 vlPFC 激活(Levy & Wagner, 2011)。在信任行为中，dlPFC 和 vlPFC 通过调节自下而上加工并对环境中的线索进行评估来减少和消除不确定性：在之前没有受托人信息的情况下，与合作性伙伴和利己主义者进行信任博弈游戏时 dlPFC 会有不同程度激活；在了解受托人情况后，当受托人出现违背信任的情况时 vlPFC 显著激活，并在违背信任的情况出现之后使个体依然选择信任对方以防止不必要的报复(Souza, Donohue, & Bunge, 2009)。

2.4 默认网络：评估关系

人们在进行信任决策时，除了考虑所处情境并做出理性分析和判断之外，往往还会对所信任的对象进行评估，这种评估是人际层面的。即使在相同情境中，面对不同的信任对象时，个体所表现出的信任程度也不尽相同。这一过程通常不需要经过严谨的推理和计算，因此往往具有感性的特点。该过程涉及的脑区被称作默认网络(Raichle et al., 2001)，它与人的社会性有关，通常在人际交往和评价他人的过程中被激活。

默认网络是信任社会认知系统的核心，负责对信任关系进行评估，其中包括颞顶联合区(TPJ)、后扣带皮层(PCC)以及背内侧前额皮层(dmPFC)等关键区域(Bressler & Menon, 2010)。TPJ 与各种社会认知功能有关，包括自我–他人区分、观点采择、他人意图推断等，这些功能使之成为在推断和归因他人意图以评估可信度时的重要区域(van Overwalle, 2009)。在以往研究中发现 TPJ 激活会随受托人年龄的增长而增加，表明该区域对他人的社交信号具有更高的敏感性(Fett et al., 2014)。PCC 能够将自下而上的注意与记忆和知觉的信息相结合，腹侧 PCC 在涉及默认网络的所有任务中激活，包括与自身相关的、与他人相关的、回忆过去、思考未来、处理概念和空间导航等任务，在信任任务中涉及对他人的评价过程与 PCC 激活显著相关(Raichle et al., 2001)。dmPFC 是自我参照加工和他人印象形成的关键区域，在各种社会任务中都有激活(Amodio & Frith, 2006)。该区域通过推断和归因他人特质来评估对方的可信度：相比与计算机进行博弈，与人类进行博弈时 dmPFC 表现出更高水平的激活。dmPFC 还会依据多次互动过程中对方的表现，以及之前传递的社会特征信息对他人的可信度进行判断(McCabe, Houser, Ryan, Smith, & Trouard, 2001)。

3 信任的认知神经网络模型

以往在神经层面对信任进行的研究更多关注信任行为发生过程中相对独立的某些脑区(如 STR、ACC、杏仁核或 AI 等)激活，而忽略了作为整体存在于大脑中的大规模脑网络的作用：与感觉、知觉和注意等初级心理过程不同，信任作为人类的高级心理过程，其激活的脑区更加广泛且复杂。同时，随着信任行为的发生和发展，其激活的脑区也会呈动态变化趋势。例如，当任务从单次博弈变成重复博弈时，可以观测到大脑的激活从 SAN (AI)转到 RWN (vSTR) (Bellucci et al., 2017)。因此，从认知神经网络的视角对信任进行探讨不仅能够在神经层面形成整合的信任模型，还可以描绘出人类互动过程中人际信任的发生、发展和转变的动态过程，并对该动态变化过程从认知神经层面给出更科学的解释。再如，与理性经济人假说预期不同，委托人更容易受社会性影响并在信任建立过程的早期表现出 DMN (dmPFC)激活，而在晚期表现出 RWN (vSTR)激活(Krueger et al., 2007)。这种脑区激活的转变反映了个体做出信任决策原因的转变：从最初没有委托人信息的情况下更多依赖社会性特征的社会驱动进行信任决策，到后来随着信任博弈游戏的进行收集到了更多委托人信息，从而导致做出信任决策的原因转变为得到更多奖励的利益驱动。为了对人际信任行为动态过程进行更全面的描述，同时也能够从更加完整的视角解释信任的发展和转变过程，本文提出了信任的认知神经网络模型(图 3)，将心理系统与大规模脑网络中的各部分进行有机整合，并形成完整的动态系统。在信任过程中心理系统中的动力系统、情感系统和认知系统起着至关重要的作用，与这三个心理系统相关的脑网络分别是奖励网络(RWN)、显著网络(SAN)、中央执行网络(CEN)和默认网络(DMN)。模型中各个成分相互作用，最终由认知系统对所有信息进行整合并做出信任或不信任的决策。与此同时，信任行为提供的反馈信息作

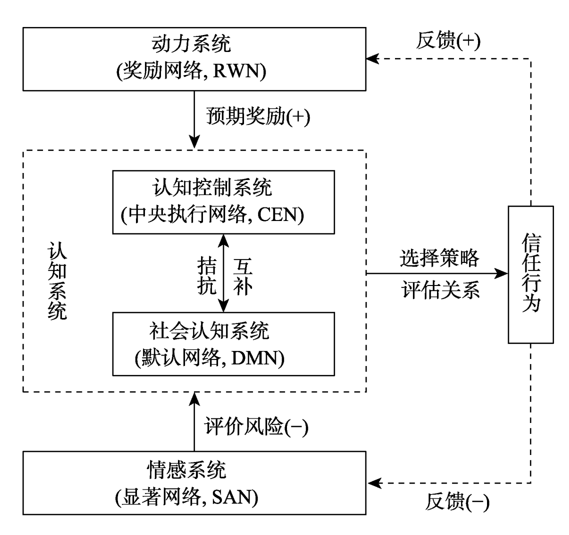


图 3 信任的认知神经网络模型

用于各个系统, 以此完善信任的神经网络模型。对该模型的理解应从以下三方面进行: (1)动力系统和情感系统对认知系统的影响; (2)认知系统内部结构及其对信任行为的影响; (3)信任行为的反馈机制。

3.1 动力系统和情感系统对认知系统的影响

心理系统中的动力系统和情感系统影响认知系统。在信任博弈游戏中, 当被试预期得到更多奖励时, RWN 显著激活, 在投资过程中表现的更加慷慨(Davidenko et al., 2018)。由于动力系统负责对奖励进行预期, 在此过程中 RWN 被激活, 认知系统会根据动力系统提供的预期结果进行判断: 动力系统预期到的奖励越多, 认知系统得到的动力也越足, 越容易做出信任的决策; 而当被试认为信任对方会有很高的风险时, SAN 激活明显, 在投资任务中会表现的更加谨慎(Uddin, 2015)。由于情感系统负责对风险进行评价, 在此过程中 SAN 被激活, 认知系统则会对情感系统提供的评价进行分析: 更高的风险往往带来更多负性情绪体验, 这会导致信任水平降低。动力系统和情感系统分别像一个“激进主义者”和一个“保守主义者”, 前者为了得到更多的奖励跃跃欲试, 后者则为了规避可能的风险畏缩不前。认知系统则是最终的“仲裁者”, 它通过对动力系统和情感系统提供的“证据”进行综合评定, 最终做出信任或者不信任的决定。由此可见, 动力系统和情感系统通过认知系统影响信任行为。

3.2 认知系统内部结构及其对信任行为的影响

认知系统包括认知控制系统和社会认知系统, 二者共同决定最终的认知结果并影响信任行为的表现。认知控制系统由 CEN 控制, 通过整合分析外界环境中各种线索以及由 RWN 和 SAN 提供的奖励和风险评估结果, 对利益得失进行精确计算, 继而做出行为决策。有证据表明 CEN 的激活与当前任务有关, 在信任博弈游戏中 CEN 的激活表明此过程涉及大量精细计算、逻辑推理等高级心理过程(Sherman et al., 2014), 因此该过程需要有充足认知资源参与, 同时也需要有充足的时间进行思维和推理, 并最终对可行性策略进行最优化选择。当外界线索有限或没有充足的认知资源进行严谨推理时, 由 CEN 控制的认知控制系统作用下降, 由 DMN 控制的社会认知系统作用凸显: 社会认知系统借由评价对方的社会性线索做出信任决策, 该过程通常不需要大量认知资源参与, 可以快速有效做出反应。此外, 一项纵向研究发现, 在 10~13 岁青少年中, 随着被试年龄的增长 PCC 和 CEN 之间的网络相关性减弱, dlPFC 与 DMN 神经隔离增加, 表明认知控制系统和社会认知系统在青春期早期就出现了结构性和功能性分离(Sherman et al., 2014)。CEN 和 DMN 在激活水平上存在显著负相关, 当有外在任务时 CEN 显著激活, 而无外在任务时 DMN 激活更明显(Raichle et al., 2001), 表明两个神经网络在功能上存在拮抗性: 某一神经网络激活时另一神经网络就会被抑制。同时, 二者又具有互补性: CEN 负责对所有可得线索进行评估并做出决策, 当认知资源有限或加工时间不足时, DMN 会根据社会性线索辅助做出信任决策。由 CEN 控制的认知控制系统和由 DMN 控制的社会认知系统共同作用于心理系统中的认知系统, 并最终影响信任行为决策。

3.3 信任行为的反馈机制

个体做出信任决策后, 可能得到积极的结果也可能得到消极的结果, 不同的结果会对个体产生不同影响, 因此信任的反馈机制尤为重要。信任行为的反馈机制作用于心理的动力系统和情感系统, 继而影响认知系统, 认知系统通过对信任结果进行评估, 调整策略并重新评估关系, 进而改变接下来的信任行为。信任过程借由这种反馈机制不断进行调整, 直至形成稳定的信任模式。在重复信任博弈游戏中, 当对方表现出合作行为时, 被试 RWN 激活程度提高, 并在接下来的投资中倾向投入更多资金(Ikemoto, 2010)。即当信任行

为得到积极结果时会出现正反馈，该反馈机制作用于动力系统，使被试对奖励的预期升高，继而通过认知系统评估，做出增加信任的决策；当对方表现出背叛行为时，被试 SAN 激活，在接下来的投资中倾向投入更少的资金(Koscik & Tranel, 2011)。即当信任行为得到消极结果时会出现负反馈，该反馈机制作用于情感系统，使被试评估到风险增加，为了避免更多的损失，认知系统会做出减少信任的决策。由此可见，信任行为的反馈机制通过正负两条反馈路线分别作用于动力系统和情感系统，进而影响认知评价过程，从而实现对信任行为的调整和优化。作为信任模型的重要组成部分，反馈机制在完善信任模型中起到重要作用，当反馈机制受到破坏，个体会表现出不正常的信任行为。信任的反馈机制是信任的认知神经网络模型中必不可少的一环，在整个信任过程中发挥着重要作用。

4 总结与展望

4.1 总结

信任的认知神经网络模型从认知神经网络视角对信任行为进行解释，能够加深我们对信任行为的理解：在心理层面上，外部环境分别作用于心理系统中的动力系统、情感系统和认知系统，认知系统在综合分析动力系统和情感系统的预期和评价之后，对信任行为做出决策；在认知神经层面上，RWN、SAN、CEN 和 DMN 四个神经网络分别发挥各自的作用，各神经网络的激活分别对应信任过程中的预期奖励、评价风险、选择策略和评估关系，不同的神经网络激活模式最终决定信任的行为结果。与此同时，信任行为的反馈机制又会反作用于决策过程，并对信任行为进行优化调整。本文在总结前人研究的基础上，以认知神经网络的视角对信任行为进行剖析，使我们对信任的理解由点上升到面，并形成网络，通过分析其内部结构和反馈机制，使我们对信任行为有更加全面和深刻的理解。

4.2 展望

人类社会中信任的发展是独一无二的。大量信任博弈的理论和实践研究为我们理解实验室情境下人际信任的神经心理机制提供可能，同时也取得了卓越的成绩。随着信任的跨学科研究技术的成熟，学者们也能更清晰的解释信任的神经心理学机制。本文提出的认知神经网络模型可以促进人们对人际信任的理解，该模型不但总结了前人研究成果，还为今后人际信任、群际信任和跨文化信任领域的研究指明方向。然而，目前的研究方法和内容依然存在局限性，今后有关信任神经机制的研究应从以下三个方面着手：

4.2.1 对人际信任过程中认知神经网络的动态变化过程进行探讨

在人际交往过程中，信任并不是一成不变的，而是会随时间和环境的变化而改变。目前，在信任领域研究中面临的最大挑战是在实验室环境中进行的神经生理学研究往往只关注信任行为出现时大脑各部位的激活状态，分析和比较大脑特定区域在不同信任决策中表现出的不同激活模式，而忽视了人际信任过程中大脑不同区域的内部联系及动态变化过程。信任的认知神经网络模型则强调各神经网络间的相互影响与联系，同时由于反馈机制的存在，更加注重信任过程中信任行为与神经网络间的动态变化过程。在认知神经网络模型中强调信任过程是动态变化的，各神经网络间也存在着动态平衡。因此，在该模型基础上，今后的研究应采取多种研究范式、多种研究手段相结合的研究方法(Hahn et al., 2014)对人际信任过程中认知神经网络的动态变化过程进行探讨。如采用事件相关电位(ERPs)与功能性核磁共振(fMRI)相结合的方式同时考察信任行为引起的神经网络在时间和空间上的动态变化过程；此外，还可以采用功能性近红外光谱技术(fNIRS)考察信任行为发展变化过程中大脑各神经网络随之出现的内部联系及动态变化过程。目前，虽然有研究对不同神经网络间激活状态的转变进行分析(Bellucci et al., 2017; Krueger et al., 2007)，但研究本身并没有聚焦神经网络，而是依然集中在对特定的脑区(如 AI、vSTR 等)激活的记录和分析。由于认知神经网络整体的动态变化更能够解释信任行为发生和转变的原因，因此对信任的认知神经网络动态变化过程进行探讨能够更深入地揭示信任的神经生理机制。

4.2.2 对信任的认知神经网络与神经递质和激素的关系进行解释

认知神经网络影响着个体的信任行为，而神经递质和激素也影响着认知神经系统，因此使用更加科学严谨的方法来解释神经递质、激素和大

脑结构在信任中的作用有助于我们更深入的理解和解释信任行为。上文中提到的中脑边缘通路、中脑皮质通路和黑质纹状体通路在神经传导过程中以多巴胺为其神经递质(Ikemoto, 2010), 该递质在神经传导过程中起到兴奋神经元的作用, 通过兴奋神经系统进而促进信任行为的发生; 此外, 有研究发现不同激素对信任行为有不同影响: 催产素水平高往往预示着更多的信任行为出现, 而睾丸酮水平高的个体则会更多表现出不信任与攻击行为(Bosch et al., 2015)。目前对神经递质和激素如何影响信任行为发生的原因尚不明确, 但神经递质能够直接作用于神经系统, 而激素也能够通过神经——体液调节间接影响神经系统。因此, 神经递质和激素对信任行为的影响可能通过认知神经系统发挥作用, 对信任的认识神经网络与神经递质和激素关系的研究有助于从更微观的角度揭示信任的神经生理机制。

4.2.3　以特殊人群为研究对象在认识神经层面对信任行为进行研究

以往对信任的研究往往采用正常成年人为研究对象, 在此基础上逐渐发展出信任的相关理论, 各种模型也不断完善, 同时取得了丰硕成果和长足进步, 信任的认知神经网络模型也在此基础上提出。随着研究的不断深入和理论的不断完善, 以在校大学生和正常成年人为被试的研究可能会遇到瓶颈, 而以特殊人群为研究对象, 如对脑损伤病人进行有关信任的研究不但可以更好的佐证认知神经研究得到的重要结论, 同时还可以对认知神经网络模型进行验证, 进而对原模型中的不足之处进行校正和完善。此外, 对于一些心理疾病(精神分裂症或边缘型人格障碍)患者来说, 发展和保持对他人信任的能力受损, 往往很难信任其他人。因此, 今后的研究也可以以精神障碍患者为研究对象, 从而揭示信任的神经心理学基础, 并为疾病诊断和新治疗方法的可行性提供客观的生物学指标和参考(Unoka, Seres, Áspán, Bódi, & Kéri, 2009)。总之, 以特殊人群为研究对象在认知神经层面对信任行为进行研究, 不但能够在理论层面上完善现有假设和模型, 还可以在实践领域中为神经障碍疾病的诊断提供依据。

参考文献

张宁, 张雨青, 吴坎坎. (2011). 信任的心理和神经生理机制. *心理科学, 34*(5), 1137–1143.

张蔚, 张振, 高宇, 段华平, 吴兴南. (2016). 经济决策中人际信任博弈的理论模型与脑机制. *心理科学进展, 24*(11), 1780–1791.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), 268–277.

Anderhub, V., Engelmann, D., & Güth, W. (2002). An experimental study of the repeated trust game with incomplete information. *Journal of Economic Behavior and Organization, 48*(2), 197–216.

Arrow, K. J. (1972). Gifts and exchanges. *Philosophy and Public Affairs, 1*(4), 343–362.

Belfi, A. M., Koscik, T. R., & Tranel, D. (2015). Damage to the insula is associated with abnormal interpersonal trust. *Neuropsychologia, 71*, 165–172.

Bellucci, G., Chernyak, S. V., Goodyear, K., Eickhoff, S. B., & Krueger, F. (2017). Neural signatures of trust in reciprocity: A coordinate-based meta-analysis. *Human Brain Mapping, 38*(3), 1233–1248.

Bellucci, G., Hahn, T., Deshpande, G., Krueger, F. (2019). Functional connectivity of specific resting-state networks predicts trust and reciprocity in the trust game. *Cognitive, Affective, and Behavioral Neuroscience, 19*(1), 165–176.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122–142.

Bosch, O. G., Eisenegger, C., Gertsch, J., von Rotz, R., Dornbierer, D., Gachet, M. S., … Quednow, B. B. (2015). Gamma-hydroxybutyrate enhances mood and prosocial behavior without affecting plasma oxytocin and testosterone. *Psychoneuroendocrinology, 62*, 1–10.

Bressler, S. L., & Menon, V. (2010). Large-scale brain networks in cognition: Emerging methods and principles. *Trends in Cognitive Sciences, 14*(6), 277–290.

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences, 4*(6), 215–222.

Clark, A. K., & Eisenstein, M. A. (2013). Interpersonal trust: An age-period-cohort analysis revisited. *Social Science Research, 42*(2), 361–375.

Davidenko, O., Bonny, J.-M., Morrot, G., Jean, B., Claise, B., Benmoussa, A., … Darcel, N. (2018). Differences in BOLD responses in brain reward network reflect the tendency to assimilate a surprising flavor stimulus to an expected stimulus. *NeuroImage, 183*, 37–46.

Declerck, C. H., Boone, C., & Emonds, G. (2013). When do people cooperate? The neuroeconomics of prosocial decision making. *Brain and Cognition, 81*(1), 95–117.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005).

Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience, 8*(11), 1611–1618.

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience, 19*(9), 1508–1519.

Engle-Warnick, J., & Slonim, R. L. (2004). The evolution of strategies in a repeated trust game. *Journal of Economic Behavior and Organization, 55*(4), 553–573.

Etkin, A., Egner, T., & Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in Cognitive Sciences, 15*(2), 85–93.

Fang, C., & Wu, B. (2019). Socially-maximal nash equilibrium distributions in large distributional games. *Economics Letters, 175*, 40–42.

Fett, A.-K. J., Gromann, P. M., Giampietro, V., Shergill, S. S., & Krabbendam, L. (2012). Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Social Cognitive and Affective Neuroscience, 9*(4), 395–402.

Fett, A.-K. J., Shergill, S. S., Gromann, P. M., Dumontheil, I., Blakemore, S.-J., Yakub, F., & Krabbendam, L. (2014). Trust and social reciprocity in adolescence-a matter of perspective-taking. *Journal of Adolescence, 37*(2), 175–184.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Hahn, T., Notebaert, K., Anderl, C., Teckentrup, V., Kaßecker, A., & Windmann, S. (2014). How to trust a perfect stranger: Predicting initial trust behavior from resting-state brain-electrical connectivity. *Social Cognitive and Affective Neuroscience, 10*(6), 809–813.

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science, 324*(5927), 646–648.

Ikemoto, S. (2010). Brain reward circuitry beyond the mesolimbic dopamine system: A neurobiological theory. *Neuroscience and Biobehavioral Reviews, 35*(2), 129–150.

Irwin, K., Edwards, K., & Tamburello, J. A. (2015). Gender, trust and cooperation in environmental social dilemmas. *Social Science Research, 50*, 328–342.

Joel, D., & Weiner, I. (2000). Striatal contention scheduling and the split circuit scheme of basal ganglia-thalamocortical circuitry: From anatomy to behaviour. *Brain Dynamics and the Striatal Complex, 12*, 209–236.

Knoch, D., & Fehr, E. (2007). Resisting the power of temptations: The right prefrontal cortex and self-control. *Annals of the New York Academy of Sciences, 1104*(1), 123–134.

Koscik, T. R., & Tranel, D. (2011). The human amygdala is necessary for developing and expressing normal interpersonal trust. *Neuropsychologia, 49*(4), 602–611.

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., … Grafman, J. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences, 104*, 20084–20089.

Krueger, F., & Meyer-Lindenberg, A. (2018). Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends in Neurosciences, 42*(2), 92–101.

Levy, B. J., & Wagner, A. D. (2011). Cognitive control and right ventrolateral prefrontal cortex: Reflexive reorienting, motor inhibition, and action updating. *Annals of the New York Academy of Sciences, 1224*(1), 40–62.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences, 98*(20), 11832–11835.

Menon, V. (2015). *Salience network*. In A. W. Toga (Ed.), Brain mapping (pp. 597–611). Waltham: Academic Press.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*(1), 167–202.

Moretto, G., Sellitto, M., & di Pellegrino, G. (2013). Investment and repayment in a trust game after ventromedial prefrontal damage. *Frontiers in Human Neuroscience, 7*, 593.

Namkung, H., Kim, S.-H., & Sawa, A. (2017). The insula: an underestimated brain area in clinical neuroscience, psychiatry, and neurology. *Trends in Neurosciences, 40*(4), 200–207.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences, 98*(2), 676–682.

Rothstein, B., & Uslaner, E. M. (2005). All for all: Equality, corruption, and social trust. *World Politics, 58*(1), 41–72.

Sherman, L. E., Rudie, J. D., Pfeifer, J. H., Masten, C. L., McNealy, K., & Dapretto, M. (2014). Development of the default mode and central executive networks across early adolescence: A longitudinal study. *Developmental Cognitive Neuroscience, 10*, 148–159.

Smith, A., & Skinner, A. S. (2000). *The wealth of nations: Books IV–V*. London: Penguin Classics.

Souza, M. J., Donohue, S. E., & Bunge, S. A. (2009). Controlled retrieval and selection of action-relevant knowledge mediated by partially overlapping regions in left ventrolateral prefrontal cortex. *NeuroImage, 46*(1), 299–307.

Stets, J. E., & Fares, P. (2019). The effects of race/ethnicity and racial/ethnic identification on general trust. *Social Science Research, 80*, 1–14.

Tzieropoulos, H. (2013). The trust game in neuroscience: A short review. *Social Neuroscience, 8*(5), 407–416.

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nature Reviews Neuroscience, 16*(1), 55–61.

Unoka, Z., Seres, I., Áspán, N., Bódi, N., & Kéri, S. (2009). Trust game reveals restricted interpersonal transactions in patients with borderline personality disorder. *Journal of Personality Disorders, 23*(4), 399–409.

van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*(3), 829–858.

The cognitive neural network model of trust

CHEN Ying; XU Minxia; WANG Xinjian

(*Zhou Enlai School of Government, Nankai University, Tianjin* 300350, *China*)

Abstract: Trust encompasses one’s willingness to take certain risks based on the expectation of the other's behavior in order to obtain positive results in the future. Previous studies focused on specific brain regions, rather than the overall activity of large-scale brain networks in trust behavior. Trust behavior is associated with the activation of multiple regions of the brain that involves the cognitive neural network. In the Cognitive Neural Network Model, trust behavior is a representation of the interaction between the motivation system, affective system and cognition system, corresponding to the activation of the reward network, salience network, central-executive network and default-mode network. The model clarifies the correspondence between psychological systems and neural networks, and explains the psychological and neural mechanisms of trust from the perspective of neuroscience. In addition, the model also emphasizes the feedback mechanism of trust behavior, yielding a complete Cognitive Neural Network Model.

Key words: trust; cognitive neural network; motivation system; affective system; cognition system